Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Manifold Learning for Jointly Modeling Topic and Visualization

Tuan M. V. Le and **Hady W. Lauw**
School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore 178902
{vmtle.2012@phdis.smu.edu.sg, hadywlauw@smu.edu.sg}

Abstract

Classical approaches to visualization directly reduce a document's high-dimensional representation into visualizable two or three dimensions, using techniques such as multidimensional scaling. More recent approaches consider an intermediate representation in topic space, between word space and visualization space, which preserves the semantics by topic modeling. We call the latter semantic visualization problem, as it seeks to jointly model topic and visualization. While previous approaches aim to preserve the global consistency, they do not consider the local consistency in terms of the intrinsic geometric structure of the document manifold. We therefore propose an unsupervised probabilistic model, called SEMAFORE, which aims to preserve the manifold in the lower-dimensional spaces. Comprehensive experiments on several real-life text datasets of news articles and web pages show that SEMAFORE significantly outperforms the state-of-the-art baselines on objective evaluation metrics.

Introduction

Visualization of high-dimensional data is an important exploratory data analysis task, which is actively studied by various academic communities. While the HCI community is interested in the presentation of information, as well as other interface aspects (Chi 2000), the machine learning community (as in this paper) is interested in the quality of dimensionality reduction (Van der Maaten and Hinton 2008), i.e., how to transform the high-dimensional representation into a lower-dimensional representation that can be shown on a scatterplot. This visualization form is simple, and widely applicable across various domains. One pioneering technique is Multidimensional Scaling (MDS) (Kruskal 1964). The goal is to preserve the *distances* in the high-dimensional space in the low-dimensional embedding. This goal also allows an objective evaluation, by verifying how well the relationships among data points are preserved by the scatterplot.

Consider the problem of visualizing documents on a scatterplot. Commonly, a document is represented as a bag of words, i.e., a vector of word counts. This high-dimensional representation would be reduced into coordinates on a visualizable 2D (or 3D) space. When applied to documents, a visualization technique for generic high-dimensional data, e.g., MDS, may not necessarily preserve the topical semantics. Words are often ambiguous, with issues such as *polysemy* (same word carries multiple senses), and *synonymy* (different words carry the same sense).

In text mining, the current approach to model semantics in documents in a way that can resolve some of this ambiguity is topic modeling, such as PLSA (Hofmann 1999) or LDA (Blei, Ng, and Jordan 2003). However, a topic model by itself is not designed for visualization. While one possible visualization is to plot documents' topic distributions on a simplex, a 2D visualization space could express only three topics, which is very limiting. By going from word space to topic space, topic modeling is also a form of dimensionality reduction. Given its utility in modeling document semantics, we are interested in achieving both forms of dimensionality reductions (visualization and topic modeling) together.

This coupling is a distinct task from topic modeling or visualization respectively, as it enables novel capabilities. For one thing, topic modeling helps to create a richer visualization, as we can now associate each coordinate on the visualization space with both topic and word distributions, providing semantics to the visualization space. For another, the tight integration potentially allows the visualization to serve as a way to explore and tune topic models, allowing users to introduce feedback to the model through a visual interface. These capabilities support several use case scenarios. One potential use case is a document organizer system. The visualization can help in assigning categories to documents, by showing how related documents have been labeled. Another is an augmented retrieval system. Given a query, the results may include not just relevant documents, but also other similar documents (neighbors in the visualization).

Problem Statement. We refer to the task of jointly modeling topics and visualization as *semantic visualization*. The input is a set of documents $\mathcal{D}$. For a specified number of topics Z and visualization dimensionality (assumed to be 2D, without losing any generality), the goal is to derive, for every document in $\mathcal{D}$, a latent coordinate on the visualization space, and a probability distribution over the Z topics. While we focus on documents in our description, the same approach would apply to visualization of other data types for which latent factor modeling, i.e., topic model, makes sense.

One approach to solve this problem is to undergo two-step reductions: going from word space to topic space using topic modeling, followed by going from topic space to coordinate space using visualization. This *pipeline* approach is not ideal, because the disjoint reductions could mean that errors may propagate from the first to the second reduction.

A better way is a *joint* approach that builds both reductions into a single, consistent whole that produces topic distributions and visualization coordinates simultaneously. The joint approach was attempted by PLSV (Iwata, Yamada, and Ueda 2008), which derives the latent parameters by maximizing the likelihood of observing the documents. This objective is known as *global consistency*, which is concerned with the "error" between the model and the observation.

Crucially, PLSV has not cared to meet the *local consistency* objective (Zhou et al. 2004), which is concerned with preserving the observed proximity or distances between documents. Local consistency is reminiscent of the objective in classical visualization (Kruskal 1964). This shortcoming is related to PLSV's assumption that the document space is Euclidean (a geometrically flat space), by sampling documents' coordinates independently in a Euclidean space.

The local consistency objective arises naturally from the assumption that the intrinsic geometry of the data is a low-rank, non-linear manifold within the high-dimensional space. This manifold assumption is well-accepted in the machine learning community (Lafferty and Wasserman 2007), and finds application in both supervised and unsupervised learning (Belkin and Niyogi 2003; Zhou et al. 2004; Zhu et al. 2003). Recently, there is a preponderance of evidence that manifold assumption also applies to text data in particular (Cai et al. 2008; Cai, Wang, and He 2009; Huh and Fienberg 2012). We therefore propose to incorporate this manifold assumption into a new unsupervised, semantic visualization model, which we call SEMAFORE.

Contributions. While visualization and topic modeling are, separately, well-studied problems, the interface between the two, semantic visualization, is a relatively new problem, with very few previous work. To our best knowledge, we are the first to propose incorporating manifold learning in semantic visualization, which is our first contribution. As a second contribution, to realize the manifold assumption, we propose a probabilistic model SEMAFORE, with a specific manifold learning framework for semantic visualization. Our third contribution is in describing the requisite learning algorithm to fit the parameters. Our final contribution is the evaluation of SEMAFORE's effectiveness on a series of real-life, public datasets of different languages, which shows that SEMAFORE outperforms existing baselines on a well-established and objective visualization metric.

Related Work

Classical visualization aims to preserve the high-dimensional similarities in the low-dimensional embedding. One pioneering work is multidimensional scaling (MDS) (Kruskal 1964), which uses linear distance. Isomap (Tenenbaum, De Silva, and Langford 2000) uses geodesic distance, whereas LLE (Roweis and Saul 2000) uses linear distance but only locally. These are followed by a body of probabilistic approaches (Iwata et al. 2007; Hinton and Roweis 2002; Van der Maaten and Hinton 2008; Bishop, Svensén, and Williams 1998). They are not meant for *semantic* visualization, as they do not model topics.

Semantic visualization is a new problem explored in very few works. The state-of-the-art is the joint approach PLSV (Iwata, Yamada, and Ueda 2008), which we use as a baseline. In the same paper, it is shown that PLSV outperforms the pipeline approach of PLSA (Hofmann 1999) followed by PE (Iwata et al. 2007). LDA-SOM (Millar, Peterson, and Mendenhall 2009) is another pipeline approach of LDA (Blei, Ng, and Jordan 2003) followed by SOM (Kohonen 1990), which produces a different type of visualization.

Semantic visualization refers to joint topic modeling and visualization of documents. A different task is topic visualization, where the objective is to visualize the topics themselves (Chaney and Blei 2012; Chuang, Manning, and Heer 2012; Wei et al. 2010; Gretarsson et al. 2012), in terms of dominant keywords, prevalence of topics, etc.

(Cai et al. 2008; Cai, Wang, and He 2009; Wu et al. 2012; Huh and Fienberg 2012) study manifold in the context of topic models only. The key difference is that we also need to contend with the visualization aspect, and not only topic modeling, which creates new research issues.

Semantic Visualization Model

We now describe our model, SEMAFORE, which stands for SEmantic visualization with MAniFOld REgularization.

Problem Definition

The input is a corpus of documents $\mathcal{D} = \{d_1, \ldots, d_N\}$. Every d_n is a bag of words, and w_{nm} denotes the m^{th} word in d_n. The total number of words in d_n is M_n. The objective is to learn, for each d_n, a latent distribution over Z topics $\{\mathrm{P}(z|d_n)\}_{z=1}^{Z}$. Each topic z is associated with a parameter θ_z, which is a probability distribution $\{\mathrm{P}(w|\theta_z)\}_{w \in W}$ over words in the vocabulary W. The words with highest probabilities for a given topic capture the semantic of that topic.

Unlike topic modeling, in semantic visualization, there is an *additional* objective, which is to learn, for each document d_n, its latent coordinate x_n on a low-dimensionality visualization space. Similarly, each topic z is associated with a latent coordinate ϕ_z on the visualization space. A document d_n's topic distribution is then expressed in terms of the Euclidean distance between its coordinate x_n and the different topic coordinates $\Phi = \{\phi_z\}_{z=1}^{Z}$, as shown in Equation 1. The closer is x_n to ϕ_z, the higher is the probability $\mathrm{P}(z|d_n)$.

$$\mathrm{P}(z|d_n) = \mathrm{P}(z|x_n, \Phi) = \frac{\exp(\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z'=1}^{Z} \exp(\frac{1}{2}||x_n - \phi_{z'}||^2)} \qquad (1)$$

Generative Process

We now describe the assumed generative process of documents based on both topics and visualization coordinates. Our focus in this paper is on the effects of the manifold assumption on the semantic visualization task. We figure that the clearest way to showcase these effects is to design a manifold learning framework over and above an existing generative process, such as PLSV (Iwata, Yamada, and Ueda 2008), which we review below.

The generative process is as follows:

1. For each topic $z = 1, \ldots, Z$:
   (a) Draw z's word distribution: $\theta_z \sim \mathrm{Dirichlet}(\alpha)$
   (b) Draw z's coordinate: $\phi_z \sim \mathrm{Normal}(0, \beta^{-1}I)$
2. For each document d_n, where $n = 1, \ldots, N$:
   (a) Draw d_n's coordinate: $x_n \sim \mathrm{Normal}(0, \gamma^{-1}I)$
   (b) For each word $w_{nm} \in d_n$:
      i. Draw a topic: $z \sim \mathrm{Multi}(\{\mathrm{P}(z|x_n, \Phi)\}_{z=1}^{Z})$
      ii. Draw a word: $w_{nm} \sim \mathrm{Multi}(\theta_z)$

Here, α is a Dirichlet prior, I is an identity matrix, β and γ control the variance of the Normal distributions. The parameters $\chi = \{x_n\}_{n=1}^{N}$, $\Phi = \{\phi_z\}_{z=1}^{Z}$, $\Theta = \{\theta_z\}_{z=1}^{Z}$, collectively denoted as $\Psi = \langle \chi, \Phi, \Theta \rangle$, are learned from documents $\mathcal{D}$ based on maximum a posteriori estimation. The log likelihood function is shown in Equation 2.

$$\mathcal{L}(\Psi|\mathcal{D}) = \sum_{n=1}^{N} \sum_{m=1}^{M_n} \log \sum_{z=1}^{Z} \mathrm{P}(z|x_n, \Phi)\mathrm{P}(w_{nm}|\theta_z) \tag{2}$$

Manifold Learning

In the above generative process, the document parameters are sampled independently, which may not necessarily reflect the underlying manifold. We therefore assume that when two documents d_i and d_j are close in the intrinsic geometry of the manifold Ω, then their parameters ψ_i and ψ_j are similar as well. To realize this assumption, we need to address several issues, including the representation of the manifold, and the mechanism to incorporate the manifold.

As a starting point, we consider the Laplacian Eigenmaps framework for manifold learning (Belkin and Niyogi 2003). It postulates that a low-dimensionality manifold relating N high-dimensional data points can be approximated by a $k-$nearest neighbors graph. The manifold graph contains an edge connecting two data points d_i and d_j, with the weight $\omega_{ij} = 1$, if d_i is in the set $\mathcal{N}_k(d_j)$ of the $k-$nearest neighbors of d_j, or d_j is in the set $\mathcal{N}_k(d_i)$. Otherwise, $\omega_{ij} = 0$. By definition, edges are symmetric, i.e., $\omega_{ij} = \omega_{ji}$. The collection of edge weights are collectively denoted as $\Omega = \{\omega_{ij}\}$. The edge weights are binary to isolate the effects of the manifold graph structure. More complex similarity-based weighting schemes are possible, and will be explored in the future.

$$\omega_{ij} = \begin{cases} 1, & \text{if } d_i \in \mathcal{N}_k(d_j) \text{ or } d_j \in \mathcal{N}_k(d_i) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

One effective means to incorporate a manifold structure into a learning model is through a regularization framework (Belkin, Niyogi, and Sindhwani 2006). This leads to a redesign of the log-likelihood function in Equation 2 into a new *regularized* function $\mathbf{L}$ (Equation 4), where Ψ consists of the parameters (visualization coordinates and topic distributions), and $\mathcal{D}$ and Ω are the documents and manifold.

$$\mathbf{L}(\Psi|\mathcal{D}, \Omega) = \mathcal{L}(\Psi|\mathcal{D}) - \frac{\lambda}{2} \cdot \mathcal{R}(\Psi|\Omega) \tag{4}$$

The first component $\mathcal{L}$ is the log-likelihood function in Equation 2, which reflects the global consistency between the latent parameters Ψ and the observation $\mathcal{D}$. The second component $\mathcal{R}$ is a regularization function, which reflects the local consistency between the latent parameters Ψ of neighboring documents in the manifold Ω. λ is the regularization parameter, commonly found in manifold learning algorithms (Belkin, Niyogi, and Sindhwani 2006; Cai et al. 2008; Cai, Wang, and He 2009), which controls the extent of regularization (we experiment with different λ's in experiments).

This design effectively subsumes PLSV as a special case when $\lambda = 0$, and enables us to directly showcase the effects of the manifold as the key differentiator in the model.

We now turn to the definition of the $\mathcal{R}$ function. The intuition is that the data points that are close in the high-dimensional space, should also be close in their low-rank representations, i.e., local consistency. The justification is the embedding maps approximate the Eigenmaps of the Laplace Beltrami operator, which provides an optimal embedding for the manifold. One function that satisfies this is $\mathcal{R}_+$ in Equation 5. Here, $\mathcal{F}$ is a distance function that operates on the low-rank space. Minimizing $\mathcal{R}_+$ leads to minimizing the distance $\mathcal{F}(\psi_i, \psi_j)$ between neighbors ($\omega_{ij} = 1$).

$$\mathcal{R}_+(\Psi|\Omega) = \sum_{i,j=1;i\neq j}^{N} \omega_{ij} \cdot \mathcal{F}(\psi_i, \psi_j) \tag{5}$$

The above level of local consistency is still insufficient, because it does not regulate how *non*-neighbors (i.e., $\omega_{ij} = 0$) behave. For instance, it does not prevent *non*-neighbors from having similar low-rank representations. Another valid objective in visualization is to keep *non*-neighbors apart, which is satisfied by another objective function $\mathcal{R}_-$ in Equation 6. $\mathcal{R}_-$ is minimized when two *non*-neighbors d_i and d_j (i.e., $\omega_{ij} = 0$) are distant in their low-rank representations. The addition of 1 to $\mathcal{F}$ is to prevent division-by-zero error.

$$\mathcal{R}_-(\Psi|\Omega) = \sum_{i,j=1;i\neq j}^{N} \frac{1 - \omega_{ij}}{\mathcal{F}(\psi_i, \psi_j) + 1} \tag{6}$$

We hypothesize that neither objective is effective on its own. A more complete objective would capture the spirits of both keeping neighbors close, and keeping *non*-neighbors apart. Therefore, in this paper, we propose a single function that combines Equation 5 and Equation 6 in a natural way. A suitable combination, which we propose in this paper, is summation, as shown in Equation 7.

$$\mathcal{R}_*(\Psi|\Omega) = \mathcal{R}_+(\Psi|\Omega) + \mathcal{R}_-(\Psi|\Omega) \tag{7}$$

Summation preserves the absolute magnitude of the distance, and helps to improve the visualization task by keeping *non*-neighbors separated on a visualizable Euclidean space. Taking the product is unsuitable, because it constraints the *ratio* of distances between neighbors to distances between *non*-neighbors. This may result in the crowding effect, where many documents are clustered together, because the relative ratio may be maintained, but the absolute distances on the visualization space could be too small.

Enforcing Manifold: Visualization vs. Topic Space. We turn to the definition of $\mathcal{F}(\psi_1, \psi_2)$. In classical manifold learning, there is one low-rank representative space. For semantic visualization, there are two: topic and visualization. We look into where and how to enforce the manifold.

At first glance, they seem equivalent. After all, they are representations of the same documents. However, this is not necessarily the case. Consider a simple example of two topics z_1 and z_2 with visualization coordinates $\phi_1 = (0, 0)$ and $\phi_2 = (2, 0)$ respectively. Meanwhile, there are three documents $\{d_1, d_2, d_3\}$ with coordinates $x_1 = (1, 1)$, $x_2 = (1, 1)$, and $x_3 = (1, -1)$. If two documents have the same coordinates, they will also have the same topic distributions. In this example, x_1 and x_2 are both equidistant from ϕ_1 and ϕ_2, and therefore according to Equation 1, they have the same topic distribution $\mathrm{P}(z_1|d_1) = \mathrm{P}(z_1|d_2) = 0.5$, and $\mathrm{P}(z_2|d_1) = \mathrm{P}(z_2|d_2) = 0.5$. If two documents have the same topic distributions, they may not necessarily have the same coordinates. d_3 also has the same topic distribution as d_1 and d_2, but a different coordinate. In fact, any coordinate of the form $(1, ?)$ will have the same topic distribution.

This example suggests that enforcing manifold on the topic space may not necessarily lead to having data points closer on the visualization space. We postulate that regularizing the visualization space is more effective. There are also advantages in computational efficiency to doing so, which we will describe further shortly. Therefore, we define $\mathcal{F}(\psi_i, \psi_j)$ as the squared Euclidean distance $||x_i - x_j||^2$ between the corresponding visualization coordinates.

Model Fitting

One well-accepted framework to learn model parameters using maximum a posteriori (MAP) estimation is the EM algorithm (Dempster, Laird, and Rubin 1977). For our model, the regularized conditional expectation of the complete-data log likelihood in MAP estimation with priors is:

$$\begin{aligned}\mathcal{Q}(\Psi|\hat{\Psi}) = & \sum_{n=1}^{N}\sum_{m=1}^{M_n}\sum_{z=1}^{Z}\mathrm{P}(z|n,m,\hat{\Psi})\log\left[\mathrm{P}(z|x_n,\Phi)\mathrm{P}(w_{nm}|\theta_z)\right] \\ & + \sum_{n=1}^{N}\log(\mathrm{P}(x_n)) + \sum_{z=1}^{Z}\log(\mathrm{P}(\phi_z)) + \sum_{z=1}^{Z}\log(\mathrm{P}(\theta_z)) \\ & - \frac{\lambda}{2}\mathcal{R}(\Psi|\Omega)\end{aligned}$$

$\hat{\Psi}$ is the current estimate. $\mathrm{P}(z|n, m, \hat{\Psi})$ is the class posterior probability of the n^{th} document and the m^{th} word in the current estimate. $\mathrm{P}(\theta_z)$ is a symmetric Dirichlet prior with parameter α for word probability θ_z. $\mathrm{P}(x_n)$ and $\mathrm{P}(\phi_z)$ are Gaussian priors with a zero mean and a spherical covariance for the document coordinates x_n and topic coordinates ϕ_z. We set the hyper-parameters to $\alpha = 0.01$, $\beta = 0.1N$ and $\gamma = 0.1Z$ following (Iwata, Yamada, and Ueda 2008).

In the E-step, $\mathrm{P}(z|n, m, \hat{\Psi})$ is updated as follows:

$$\mathrm{P}(z|n,m,\hat{\Psi}) = \frac{\mathrm{P}(z|\hat{x}_n,\hat{\Phi})\mathrm{P}(w_{nm}|\hat{\theta}_z)}{\sum_{z'=1}^{Z}\mathrm{P}(z'|\hat{x}_n,\hat{\Phi})\mathrm{P}(w_{nm}|\hat{\theta}_{z'})}$$

In the M-step, by maximizing $\mathcal{Q}(\Psi|\hat{\Psi})$ w.r.t θ_{zw}, the next estimate of word probability θ_{zw} is as follows:

$$\theta_{zw} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M_n}\mathrm{I}(w_{nm}=w)\mathrm{P}(z|n,m,\hat{\Psi}) + \alpha}{\sum_{w'=1}^{W}\sum_{n=1}^{N}\sum_{m=1}^{M_n}\mathrm{I}(w_{nm}=w')\mathrm{P}(z|n,m,\hat{\Psi}) + \alpha W}$$

$\mathrm{I}(.)$ is the indicator function. ϕ_z and x_n cannot be solved in a closed form, and are estimated by maximizing $\mathcal{Q}(\Psi|\hat{\Psi})$ using quasi-Newton (Liu and Nocedal 1989).

We compute the gradients of $\mathcal{Q}(\Psi|\hat{\Psi})$ w.r.t ϕ_z and x_n respectively as follows:

$$\begin{aligned}\frac{\partial Q}{\partial \phi_z} &= \sum_{n=1}^{N}\sum_{m=1}^{M_n}\left(\mathrm{P}(z|x_n,\Phi) - \mathrm{P}(z|n,m,\hat{\Psi})\right)(\phi_z - x_n) - \beta\phi_z \\ \frac{\partial Q}{\partial x_n} &= \sum_{m=1}^{M_n}\sum_{z=1}^{Z}\left(\mathrm{P}(z|x_n,\Phi) - \mathrm{P}(z|n,m,\hat{\Psi})\right)(x_n - \phi_z) - \gamma x_n \\ & \quad - \frac{\lambda}{2}\frac{\partial\mathcal{R}(\Psi|\Omega)}{\partial x_n}\end{aligned}$$

The gradient of $\mathcal{R}(\Psi|\Omega)$ w.r.t. x_n is computed as follows:

$$\begin{aligned}\frac{\partial\mathcal{R}(\Psi|\Omega)}{\partial x_n} = & \sum_{j=1;j\neq n}\left(4\omega_{nj}(x_n - x_j)\right) \\ & - \sum_{j=1;j\neq n}\left(4(1-\omega_{nj})\frac{(x_n - x_j)}{(\mathcal{F}(\psi_n,\psi_j)+1)^2}\right)\end{aligned}$$

As mentioned earlier, there is an efficiency advantage to regularizing on the visualization space. $\mathcal{R}(\Psi|\Omega)$ does not contain the variable ϕ_z if we do regularization on visualization space. The complexity of computing all $\frac{\partial\mathcal{R}(\Psi|\Omega)}{\partial x_n}$ is $O(N^2)$. In contrast, if we do regularization on topic space, we have to take the gradient of $\mathcal{R}(\Psi|\Omega)$ w.r.t to ϕ_z. That contributes towards a greater complexity of $O(Z^2 \times N^2)$ to compute all $\frac{\partial\mathcal{R}(\Psi|\Omega)}{\partial\theta_z}$. Therefore, regularization on topic space would run much slower than on visualization space.

Experiments

Experimental Setup

Datasets. We use three real-life, publicly available datasets[1] for evaluation. $20News$ contains newsgroup articles (in English) from 20 classes. $Reuters8$ contains newswire articles (in English) from 8 classes. $Cade12$ contains web pages (in Brazilian Portuguese) classified into 12 classes. These are benchmark datasets frequently used for document classification. While our task is fully unsupervised, the ground-truth class labels are useful for an objective evaluation.

Following (Iwata, Yamada, and Ueda 2008), we create balanced classes by sampling fifty documents from each class. This results in, *for one sample*, 1000 documents for $20News$, 400 for $Reuters8$, and 600 for $Cade12$. The vocabulary sizes are 5.4K for $20News$, 1.9K for $Reuters8$, 7.6K for $Cade12$. As the algorithms are probabilistic, we generate five samples for each dataset, conduct five runs for each sample, and average the results across a total of 25 runs.

[1] http://web.ist.utl.pt/acardoso/datasets/

Metric. For a suitable metric, we return to the fundamental principle that a good visualization should preserve the relationship between documents (in high-dimensional space) in the lower-dimensional visualization space. User studies, even when well-designed, could be overly subjective and may not be repeatable across different users reliably. Therefore, for a more objective evaluation, we rely on the ground-truth class labels found in the datasets. This is a well-established practice in many clustering and visualization works in machine learning. The basis for this evaluation is the reasonable assumption that documents of the same class are more related than documents of different classes, and therefore a good visualization would place documents of the same class as near neighbors on the visualization space.

For each document, we hide its true class, and predict its class by taking the majority class among its t-nearest neighbors as determined by Euclidean distance on the visualization space. $Accuracy(t)$ is defined as the fraction of documents whose predicted class matches the truth. By default, we use $t = 50$, because there are 50 documents in each class. The same metric is used in (Iwata, Yamada, and Ueda 2008). While accuracy is computed based on documents' coordinates, the same trends will be produced if computed based on topic distributions (due to their coupling in Equation 1).

Comparative Methods. As semantic visualization seeks to ensure consistency between topic model and visualization, the comparison focuses on methods producing *both* topics and visualization coordinates, which are listed in Table 1. SEMAFORE is our proposed method that incorporates manifold learning into semantic visualization. PLSV is the state-of-the-art, representing the joint approach without manifold. LDA/MDS represents the pipeline approach, topic modeling with LDA (Blei, Ng, and Jordan 2003), followed by visualizing documents' topic distributions with MDS (Kruskal 1964). There are other pipeline methods, shown inferior to PLSV in (Iwata, Yamada, and Ueda 2008), which are not reproduced here to avoid duplication.

Parameter Study

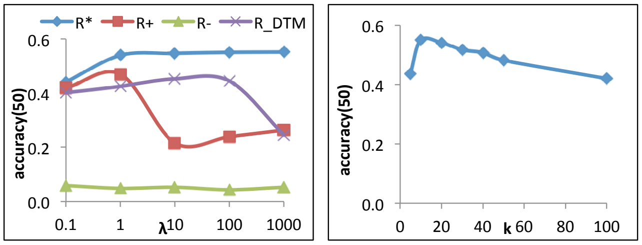
We study the effects of model parameters. Due to space constraint, we rely on $20News$ for this discussion (similar observations can be made for the other two datasets). When unvaried, the defaults are number of topics $Z = 20$, neighborhood size $k = 10$, and regularization $\mathcal{R}_*$ with $\lambda = 1$.

Regularization. One consideration is the regularization component, both the function as well as the λ. To investigate this, we compare our three proposed functions: neighbor only $\mathcal{R}_+$ (Equation 5), *non*-neighbor only $\mathcal{R}_-$ (Equation 6), and combined $\mathcal{R}_*$ (Equation 7). For completeness, we include another function $\mathcal{R}_{DTM}$, proposed by (Huh and Fienberg 2012) for a different context (topic modeling alone).

Figure 1(a) shows the accuracy at different settings of $\lambda \in [0.1, 1000]$ (log scale). Among the three proposed functions, $\mathcal{R}_*$ has the best accuracy at any λ, which is as hypothesized given that it incorporates the manifold information from both neighbors and *non*-neighbors. $\mathcal{R}_*$ is also significantly better than $\mathcal{R}_{DTM}$, which is not designed for semantic visualization. $\mathcal{R}_+$ and $\mathcal{R}_-$ are worse than $\mathcal{R}_{DTM}$, which also incorporates some information from non-neighbors. As λ increases, $\mathcal{R}_*$'s accuracy initially increases, but stabilises from $\lambda = 1$ onwards. Subsequently, we use $\mathcal{R}_*$ at $\lambda = 1$.

	Visualization	Topic model	Joint model	Manifold
SEMAFORE	X	X	X	X
PLSV	X	X	X	
LDA/MDS	X	X		

Table 1: Comparative Methods



a. Regularization

b. Neighborhood Size

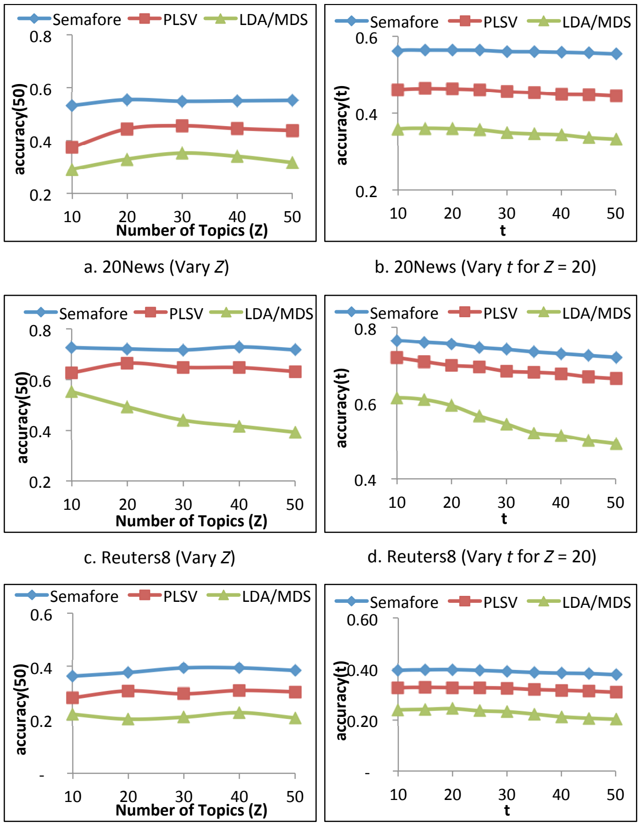
Figure 1: SEMAFORE: Vary Parameters

Neighborhood Size. To construct the manifold graph $\Omega = \{\omega_{ij}\}$, we represent each document as a tf-idf vector. We have experimented with different vector representations, including word counts and term frequencies, and found tf-idf to give the best results. The distance between two document vectors is measured using cosine distance. The $k-$nearest neighbors to i is assigned $\omega_{ij} = 1$. The rest are assigned $\omega_{ij} = 0$. In Figure 1(b), we plot the accuracy for different k's, with $\mathcal{R}_*$ and $\lambda = 1$. As k increases, the accuracy at first increases, and then decreases. This is expected as neighbors that are too far away may no longer be related, and begin to introduce noise into the manifold. The optimum is $k = 10$.

Comparison against Baselines

Accuracy. In Figure 2(a), we show the performance in $accuracy(50)$ on $20News$, while varying the number of topics Z. Figures 2(c) and 2(e) show the same for $Reuters8$ and $Cade12$ respectively. From these figures, we draw the following observations about the comparative methods. **(#1)** SEMAFORE performs the best on all datasets across various numbers of topics (Z). The margin of performance with respect to PLSV is statistically significant in all cases. SEMAFORE beats PLSV by 20% to 42% on $20News$, by 8–21% on $Reuters8$, and by 22–32% on $Cade12$. This effectively showcases the utility of manifold learning in enhancing the quality of visualization. **(#2)** PLSV performs better than LDA/MDS, which shows that there is utility to having a *joint*, instead of separate, modeling of topics and visualization. **(#3)** In Figures 2(b), 2(d), and 2(f), we show the $accuracy(t)$ at different t's for $Z = 20$ for the three datasets. The $accuracy(t)$ values are stable. At any t, the comparison shows outperformance by SEMAFORE over the baselines. **(#4)** The above accuracy results are based on visualization coordinates. We have also computed accuracies based on topic distributions, which have similar trends.

Heretofore, we will focus on the comparison between SEMAFORE and the closest competitor PLSV.



a. 20News (Vary *Z*)

b. 20News (Vary *t* for *Z* = 20)

c. Reuters8 (Vary *Z*)

d. Reuters8 (Vary *t* for *Z* = 20)

e. Cade12 (Vary *Z*)

f. Cade12 (Vary *t* for *Z* = 20)

Figure 2: Accuracy Comparison

Visualization. To provide an intuitive appreciation, we briefly describe a qualitative comparison of visualizations. Figure 4 shows a visualization of $20News$ dataset as a scatterplot (best seen in color). Each document has a coordinate, and is assigned a shape and color based on its class. Each topic also has a coordinate, drawn as a black, hollow circle. SEMAFORE's Figure 4(a) shows that the different classes are well separated. There are distinct blue cluster and purple cluster on the right for hockey and baseball classes respectively, orange and pink clusters at the top for cryptography and medicine, etc. Beyond individual classes, the visualization also places related classes[2] nearby. Computer-related classes are found on the lower left. Politics and religion classes are on the lower right. Figure 4(b) by PLSV is significantly worse. There is a lot of crowding at the center. For instance, motorcycle (green) and autos (red) are mixed at the center without a good separation.

Figure 5 shows the visualization outputs for $Reuters8$ dataset. SEMAFORE in Figure 5(a) is better at separating the eight classes into distinct clusters. In an anti-clockwise direction from the top, we have green triangles (*acq*), red squares (*crude*), purple crosses (*ship*), blue asterisks (*grain*), red dashes (*interest*), navy blue diamonds (*money-fx*), orange circles (*trade*), and finally the light blue *earn* on the upper right. In comparison, PLSV in Figure 5(b) shows that several classes are intermixed at the center, including red dashes (*interest*), orange circles (*trade*), and navy blue diamonds (*money-fx*). Figure 6 shows the visualization outputs for $Cade12$. This is the most challenging dataset. Even so, SEMAFORE in Figure 6(a) still achieves a better separation between the classes, as compared to PLSV in Figure 6(b).

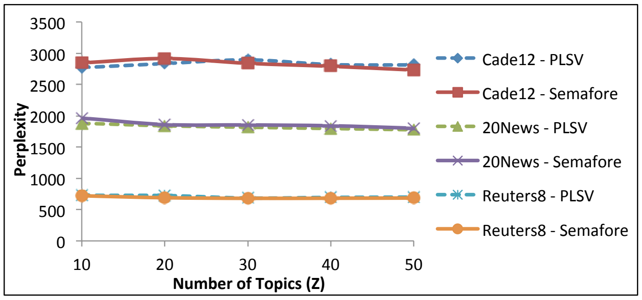


Figure 3: Perplexity Comparison

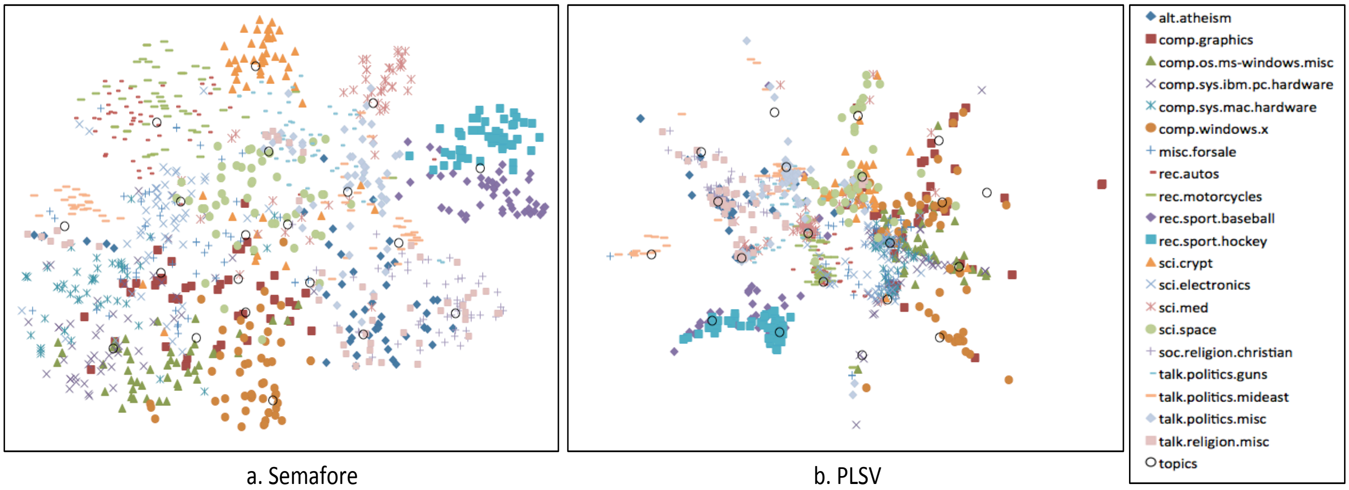
Perplexity. One question is whether SEMAFORE's gain in visualization quality over the closest baseline PLSV is at the expense of the topic model. To investigate this, we compare the *perplexity* of SEMAFORE and PLSV, which share a core generative process. Perplexity is a well-accepted metric that measures the generalization ability of a topic model on a held-out test set. For each dataset, we draw a sixth sample as test set, excluding documents that already exist in the first five samples. Perplexity is measured as $\exp\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\}$, where M is the number of documents in the test set, N_d is the number of words in a document, and $p(\mathbf{w}_d)$ is the likelihood of a test document by a topic model. Lower perplexity is better.

Figure 3 shows the perplexity as the number of topics Z varies. Perplexity values for both SEMAFORE and PLSV are close. In most cases (13 out of 15 cases), t-tests at 1% significance level indicate that the differences are not significant, except for a couple of data points (in 1 case SEMAFORE is better, in 1 case PLSV is better). This result is not unexpected, as both are optimized for log-likelihood. SEMAFORE further ensures that the document parameters (coordinates and topic distributions) that optimize the log-likelihood also better reflect the manifold. Our emphasis is on enhancing visualization, and indeed SEMAFORE's gain in visualization quality has not hurt the generalizability of its topic model.

Conclusion

We address the semantic visualization problem, which jointly conducts topic modeling and visualization of documents. We propose a new framework to incorporate manifold learning within a probabilistic semantic visualization model called SEMAFORE. Experiments on real-life datasets show that SEMAFORE significantly outperforms the baselines in terms of visualization quality, providing evidence that manifold learning, together with joint modeling of topics and visualization, is important for semantic visualization.

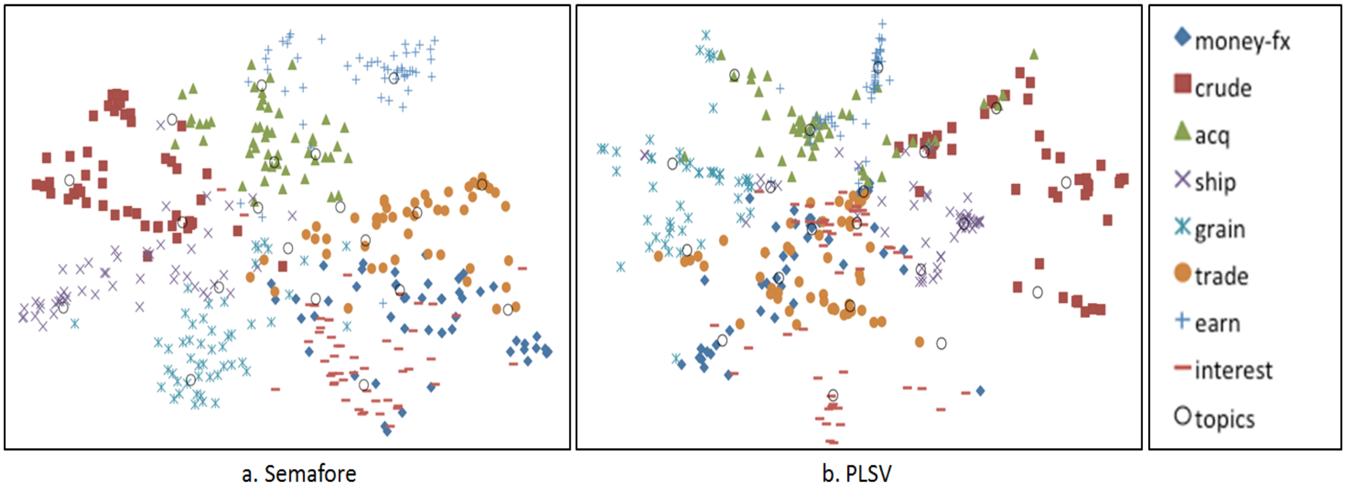
[2] http://qwone.com/∼jason/20Newsgroups/



a. Semafore

b. PLSV

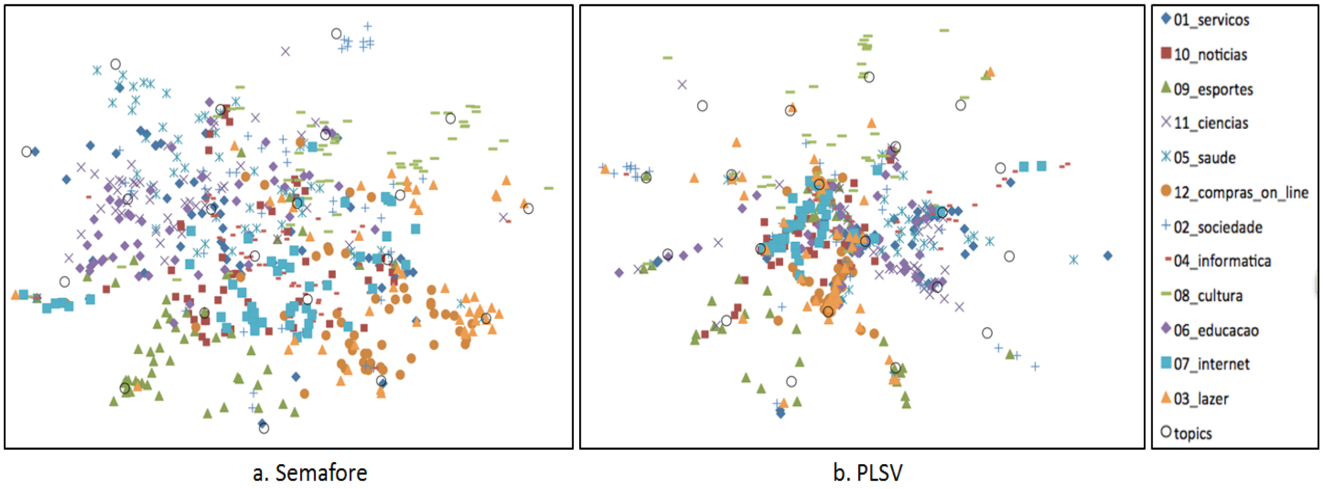
Figure 4: Visualization of $20News$ for $Z = 20$



a. Semafore

b. PLSV

Figure 5: Visualization of $Reuters8$ for $Z = 20$



a. Semafore

b. PLSV

Figure 6: Visualization of $Cade12$ for $Z = 20$

References

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15(6):1373–1396.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR* 7:2399–2434.

Bishop, C. M.; Svensén, M.; and Williams, C. K. 1998. GTM: The generative topographic mapping. *Neural Computation* 10(1):215–234.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Cai, D.; Mei, Q.; Han, J.; and Zhai, C. 2008. Modeling hidden topics on document manifold. In *CIKM*.

Cai, D.; Wang, X.; and He, X. 2009. Probabilistic dyadic data analysis with local and global consistency. In *ICML*.

Chaney, A. J.-B., and Blei, D. M. 2012. Visualizing topic models. In *ICWSM*.

Chi, E. H.-h. 2000. A taxonomy of visualization techniques using the data state reference model. In *InfoVis*, 69–75.

Chuang, J.; Manning, C. D.; and Heer, J. 2012. Termite: visualization techniques for assessing textual topic models. In *AVI*, 74–77.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38.

Gretarsson, B.; O'donovan, J.; Bostandjiev, S.; Höllerer, T.; Asuncion, A.; Newman, D.; and Smyth, P. 2012. TopicNets: Visual analysis of large text corpora with topic modeling. *TIST* 3(2):23.

Hinton, G. E., and Roweis, S. T. 2002. Stochastic neighbor embedding. In *NIPS*, 833–840.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57.

Huh, S., and Fienberg, S. E. 2012. Discriminative topic modeling based on manifold learning. *TKDD* 5(4):20.

Iwata, T.; Saito, K.; Ueda, N.; Stromsten, S.; Griffiths, T. L.; and Tenenbaum, J. B. 2007. Parametric embedding for class visualization. *Neural Computation* 19(9):2536–2556.

Iwata, T.; Yamada, T.; and Ueda, N. 2008. Probabilistic latent semantic visualization: topic model for visualizing documents. In *KDD*, 363–371.

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE* 78(9):1464–1480.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29(1):1–27.

Lafferty, J. D., and Wasserman, L. 2007. Statistical analysis of semi-supervised regression. In *NIPS*, 801–808.

Liu, D. C., and Nocedal, J. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45:503–528.

Millar, J. R.; Peterson, G. L.; and Mendenhall, M. J. 2009. Document clustering and visualization with latent dirichlet allocation and self-organizing maps. In *FLAIRS Conference*, volume 21, 69–74.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Tenenbaum, J. B.; De Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.

Van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR* 9(2579-2605):85.

Wei, F.; Liu, S.; Song, Y.; Pan, S.; Zhou, M. X.; Qian, W.; Shi, L.; Tan, L.; and Zhang, Q. 2010. Tiara: a visual exploratory text analytic system. In *KDD*, 153–162.

Wu, H.; Bu, J.; Chen, C.; Zhu, J.; Zhang, L.; Liu, H.; Wang, C.; and Cai, D. 2012. Locally discriminative topic modeling. *Pattern Recognition* 45(1):617–625.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. *NIPS* 16(16).

Zhu, X.; Ghahramani, Z.; Lafferty, J.; et al. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, 912–919.